Supplementary Information

An atlas of genetic influences on human blood metabolites

Table of Contents

Supplementary Note

Genetic and metabolomic data collection

Blood sampling

Blood samples were collected from the TwinsUK cohort after at least 6 hours of fasting predominantly overnight. For plasma EDTA storage, blood on K2 EDTA was collected. Bloods were centrifuged for 10 minutes at 3,000RPM and plasma was removed from the tubes as the top, yellow, clear layer of liquid. Aliquoting of specimens was in 1.5 ml skirted microcentrifuge tubes. All tubes were filled approximately up to 0.5 ml plasma. Until the analysis, samples were stored in freezers at -80C. Blood samples in KORA F4 were collected between 2006 and 2008, as part of the follow-up examination. To avoid variation due to circadian rhythm, blood was drawn in the morning between 8:00 am and 10:30 am after at least 10 hours overnight fasting. Material was drawn into serum gel tubes, gently inverted twice, kept for 30 min at room temperature (18−25°C) to obtain complete coagulation, and centrifuged for 10 min at 2,750*g* (at 15°C). Serum was divided into aliquots and kept for a maximum of 6 hours at 4°C, after which it was deep frozen to −80°C until analysis. Differences in fasting time between the two studies could influence variation in the metabolite levels [1]. This, however, does not affect the validity of the metabolite/SNP associations that replicate over both cohorts in the present study.

Metabolomic data acquisition and pre-processing

Sample Preparation for Glabal Metabolomics. Samples were stored at –70°C until processed. Sample preparation was carried out as described previously [2] at Metabolon, Inc. Briefly, recovery standards were added prior to the first step in the extraction process for quality control purposes. To remove protein, dissociate small molecules bound to protein or trapped in the precipitated protein matrix, and to recover chemically diverse metabolites, proteins were precipitated with methanol under vigorous shaking for 2 min (Glen Mills Genogrinder 2000) followed by centrifugation. The resulting extract was divided into four fractions: one for analysis by ultra high performance liquid chromatography-tandem mass spectrometry (UPLC-MS/MS; positive mode), one for analysis by UPLC-MS/MS (negative mode), one for analysis by gas chromatography–mass spectrometry (GC-MS), and one sample was reserved for backup.

Three types of controls were analyzed in concert with the experimental samples: samples generated from a pool of human plasma (extensively characterized by Metabolon, Inc.) served as technical replicate throughout the data set; extracted water samples served as process blanks; and a cocktail of standards spiked into every analyzed sample allowed instrument performance monitoring. Instrument variability was determined by calculating the median relative standard deviation (RSD) for the standards that were added to each sample prior to injection into the mass spectrometers (median RSD=5%; n=30 standards). Overall process variability was determined by calculating the median RSD for all endogenous metabolites (i.e., non-instrument standards) present in 100% of the pooled human plasma samples (median RSD=16.7%; n=490 metabolites). Experimental samples and controls were randomized across the platform run.

Mass Spectrometry Analysis. Non-targeted MS analysis was performed at Metabolon, Inc. Extracts were subjected to either GC-MS [3] or UPLC-MS/MS [2]. The chromatography was standardized and, once the method was validated no further changes were made. As part of Metabolon's general practice, all columns were purchased from a single manufacturer's lot at the outset of experiments. All solvents were similarly purchased in bulk from a single manufacturer's lot in sufficient quantity to complete all related experiments. For each sample, vacuum-dried samples were dissolved in injection solvent containing eight or more injection standards at fixed concentrations, depending on the platform. The internal standards were used both to assure injection and chromatographic consistency. Instruments were tuned and calibrated for mass resolution and mass accuracy daily.

The UPLC-MS/MS platform utilized a Waters Acquity UPLC and a ThermoFisher LTQ mass spectrometer, which included an electrospray ionization source and a linear ion-trap mass analyzer operated at nominal mass resolution. The instrumentation was set to monitor for positive ions in acidic extracts or negative ions in basic extracts through independent injections. Extracts were reconstituted, loaded onto columns (Waters UPLC BEH C18-2.1×100 mm, 1.7 μm), and gradient-eluted with water and 95% methanol containing 0.1% formic acid (acidic extracts) or 6.5 mM ammonium bicarbonate (basic extracts). Columns were washed and reconditioned after every injection. The instrument was set to scan 99–1000 m/z and alternated between MS and data-dependent MS^2 scans using dynamic exclusion. The scan speed was approximately six scans per s (three MS and three MS/MS scans).

The samples destined for analysis by GC-MS were dried under vacuum desiccation for a minimum of 18 h prior to being derivatized under dried nitrogen using bistrimethyl-silyltrifluoroacetamide. Derivatized samples were separated on a 5% phenyldimethyl silicone column with helium as carrier gas and a temperature ramp from 60° to 340°C within a 17-min period. All samples were analyzed on a Thermo-Finnigan Trace DSQ MS operated at unit mass resolving power with electron impact ionization and a 50–750 atomic mass unit scan range.

Compound Identification, Quantification and Data Curation. Metabolites were identified by automated comparison of the ion features in the experimental samples to a reference library of chemical standard entries that included retention time, molecular weight (m/z), preferred adducts, and in-source fragments as well as associated MS spectra and curated by visual inspection for quality control using software developed at Metabolon [4]. Identification of known chemical entities is based on comparison to metabolomic library entries of purified standards. Over 4,000 commercially available purified standard compounds have been acquired and registered into LIMS for distribution to both the LC/MS and GC/MS platforms for determination of their detectable characteristics. An additional 5,300 mass spectral entries have been created for structurally unnamed biochemicals, which have been identified by virtue of their recurrent nature (both chromatographic and mass spectral). These compounds have the potential to be identified by future acquisition of a matching purified standard or by classical structural analysis. Peaks were quantified using area-under-the-curve. Raw area counts for each metabolite in each sample were normalized to correct for variation resulting from instrument inter-day tuning differences by the median value for each run-day, therefore, setting the medians to 1.0 for each run. This preserved variation between samples but allowed metabolites of widely different raw peak areas to be compared on a similar graphical scale. Missing values were imputed with the observed minimum after normalization.

Metabolomic data overview and processing

Datasets. A total of 529 different metabolites were measured in this study on human plasma from 5,004 individuals of the TwinsUK cohort. Data for 1,052 additional TwinsUK participants was generated previously using the same analytical platform and protocols, and was merged to the current dataset. A subset of this latter dataset (258 metabolites) was reported in [5]. Data for 1,768 KORA F4 participants was measured previously to this study [5,6]. The merged final dataset included a total of 503 metabolites measured in 6,056 TwinsUK samples and 517 metabolites measured in 1,768 KORA S4 samples, of which 486 overlap between two cohorts (**Supplementary Table 1**).

Metabolite Identity. Out of the 529 total metabolites, 310 were classified as known, meaning that their analytical characteristics (specific retention time, one or multiple masses (e.g. from adducts), and the fragmentation pattern of the primary ion(s)) match the characteristics of a metabolite with known chemical structure in Metabolon's spectra library [7,8]. The known metabolites quantified in this study are spanning a wide range of relevant biochemical classes (amino acids, acylcarnitines, sphingomyelins, glycerophospholipids, carbohydrates, vitamins, lipids, nucleotides, peptides, xenobiotics and steroids; a full list of metabolites is given in **Supplementary Table 2**). A total of 219 'unknown' compounds were also measured. These unknowns correspond to metabolites in the library, whose chemical identity had not yet been definitively elucidated at the time of analysis [6].

Quality Control. Patterns of missingness for each sample and for each metabolite were investigated, and one TwinsUK sample with high missing rate (83%) was excluded. No metabolite was excluded because of data missingness. For the remaining samples, the correlation between metabolite missingness rates and experimental batches (i.e. run-days 1-27, 28-49, 50-71, 72-97, 98-122 and 123-147) was assessed. Due to calibration of the machines at periodical time points based on the date on which output files from Metabolon were generated, the missingness rate was shown to be correlated with the influence of experimental batches. Experimental batch effect was added as a covariate in association analysis, and a data normalization step was applied to adjust for variation due to instrument run-day tuning differences. For each metabolite, the raw value was corrected within the same run-day by registering the run-day medians to equal to one and normalizing each data point proportionately. A log transformation with base 10 was applied to all the metabolites, following previous work [5]. After transformation, data points laying more than 4 standard deviations from the mean of each metabolite concentration were excluded. The number of samples, minimum and maximum values, mean and standard deviation for each metabolite in the final QC-ed data are reported in **Supplementary Table 2**.

Genotyping and HapMap2 imputation

The genotyping and imputation steps for the TwinsUK and KORA F4 cohorts have been described previously in detail and are briefly described here [9-11].

TwinsUK. Genotyping. Genotyping of the TwinsUK dataset was done with a combination of Illumina arrays (HumanHap300, HumanHap610Q, 1M-Duo and 1.2MDuo). Normalised intensity data for each of the three arrays were processed using the Illluminus calling algorithm. No calls were assigned if an individual's most likely genotyped was called with less than a posterior probability threshold of 0.95. Finally, intensity cluster plots of significant SNPs were visually inspected for over-dispersion biased no calling, and/or erroneous genotype assignment. SNPs exhibiting any of these characteristics were discarded. **Data QC**. Similar exclusion criteria were applied to each of the three datasets separately. *Samples*: Exclusion criteria were: (i) sample call rate <98%, (ii) heterozygosity across all SNPs ≥2 s.d. from the sample mean; (iii) evidence of non-European ancestry as assessed by PCA comparison with HapMap3 populations; (iv) observed pairwise IBD probabilities suggestive of sample identity errors; (v). We corrected misclassified monozygotic and dizygotic twins based on IBD probabilities. *SNPs*. Exclusion criteria were (i) Hardy-Weinberg p-value$<10^{-6}$, assessed in a set of unrelated samples; (ii) MAF<1%, assessed in a set of unrelated samples; (iii) SNP call rate <97% (SNPs with MAF≥5%) or < 99% (for 1%≤MAF<5%). Alleles of all three datasets were aligned to the Human Genome (Build36) forward strand. **Data merge**. Data from the three genotyping panels were merged to generate a single dataset for imputation. Prior to merging, strict quality control was carried out on each pairwise dataset to exclude SNPs and samples showing evidence for genotyping bias in any two dataset. Quality criteria were as follows: (i) concordance for samples typed in two different datasets was set at >99%; (ii) concordance for duplicate SNPs >99%; (iii) Hardy-Weinberg p-value$<10^{-6}$, assessed in a set of unrelated samples; (iv) observed pairwise IBD probabilities suggestive of sample identity errors. Furthermore, systematic genotyping bias between any two datasets was assessed by carrying out logistic regression after randomly assigning case status to one of the two datasets. No inflation of summary statistics was observed. After quality control, the three datasets were merged, and duplicate individuals removed. The merged dataset consists of 5,654 individuals (2,040 from the HumanHap300, 3,461 from the HumanHap610Q and 153 from the HumanHap1M and 1.2MDuo arrays) and a variable number of SNPs depending on the

SNP array used (HumanHap300: 303,940, HumanHap610Q: 553,487, HumanHap1M and 1.2MDuo: 874,733). The HumanHap1M and 1.2MDuo datasets were further reduced to the SNP content of the HumanHap610Q array (553,487 SNPs) for imputation. **Imputation**. Imputation was performed using the IMPUTE software package (v2) [12] using a stepwise procedure. First, the sparser HumanHap300 dataset was imputed to the HumanHap610Q content using phased TwinsUK HumanHap610Q haplotypes as reference. Subsequently, the combined panel was imputed using reference haplotypes from the HapMap2 project (rel 22, combined CEU+YRI+ASN panels).

KORA F4. Genotyping of the KORA F4 population was carried out using the Affymetrix GeneChip array 6.0 and the genotypes were determined using Birdseed2 clustering algorithm. The criteria of call rate > 95% and p(Hardy-Weinberg) > 10^{-6} were applied as filters for SNP quality: 655,658 autosomal SNPs satisfied these criteria. Imputation was done using IMPUTE v0.4.2 [12] based on HapMap 2 (**Supplementary Table 1**).

Interpretation and reporting of ratios

The following rules were then applied for reporting an association with ratio in the manuscript and supplementary online website. Evidence for ratios was reported only when statistical evidence for the ratio was stronger than for a metabolite alone. Namely, ratios were considered only if their P-gain > 250, where P-gain = min(P(Metabolite A), P(Metabolite B))/P(A/B), following the formalization previously presented [13].

- If a locus was associated more strongly with a ratio than with a metabolite (as indicated by P-gain>250), both the metabolite and the best ratio of all significant ratios were reported;
- If a locus was associated with both a metabolite and one or more ratio at genome-wide significance, but P-gain<250, only the association with the metabolite was reported;
- If a locus was associated only with one or more ratios, the ratios were reported. Only 8 of 145 associations were explained uniquely by a ratio, and namely *ALPL, F12, PRRC2A, DDC, AKR1C4, SLC27A2, GOT2* and *APOE*.

We applied less stringent criteria for reporting metabolites and ratios into the metabolomics GWAS database to enable retrieval of suggestive associations for metabolite-focused or locus-focused analyses of interest. The criteria for inclusion in the GWAS database were:

- For metabolites: all associations with P-value of 10^{-4}
- For ratios: all associations with $P<1.00x10^{-6}$; P-gain≥10

Among the associations with ratios thus reported, it is worth noting that multiple statistical and biological interpretations are possible. A subset of the associations may be interpreted as the SNP affecting a biochemical reaction where one molecule links to the substrate and one to the product (classed as 'Activity' in the column Biochemical locus-metabolite relationship [PMID] in **Supplementary Table 6**). Other ratios exist where both molecules are linked to a substrate or both linked to a product, and where presumably the effect of the genetic variant is to cause one molecule to be consumed or acted on faster than the other ('Selectivity'). Other more can be interpreted as 'Normalizing' ratios, where one metabolite is probably (related to) a substrate or product and the other is not, but they are both in the same class (e.g., "amino acids"). Finally, 'Unknowns' ratios include all cases where one or both molecules are still unknowns or where there is no candidate causal gene (all others).

Heritability of metabolites explained by metabolic loci

Heritability analyses applied to metabolite data allow estimating the relative contribution of genetic and non-genetic factors to metabolite variance, and the extent to which genetic factors contribute to variation in metabolite concentrations. We applied the traditional twin (ACE) model to the TwinsUK dataset to partition metabolite variance into their additive genetic, shared environmental and unique environmental contributions [14]. The median metabolite ACE-heritability was 0.25 (IQR=0.14-0.35), with the highest heritabilities estimated for several lipid and steroid derivatives (for instance butyrylcarnitine, $h^2=0.76$ or androsterone sulfate, $h^2=0.71$), remarkably the tobacco metabolite cotinine ($h^2=0.75$), the glycemic marker 1,5-anhydroglucitol ($h^2=0.61$) and several unknown compounds (**Figure 3** and **Supplementary Table 7**).

The fraction of heritability explained by known loci was then estimated as the ratio of total variance of the metabolite to the variance explained by the regression model including all lead SNPs associated with a given metabolite. The fraction of heritability explained by the discovered loci was high (median 6.9%, range 1-62%). SNPs explained more than 50% heritability in the case of four metabolites (5-oxoproline, X-12092, Nacetylornithine, butyrylcarnitine), and more than 20% heritability in the case of 31 metabolites. These results suggest a high individual contribution of SNPs to metabolite variance. These estimates tend to be orders of magnitude greater compared to derived but more complex endpoints such as HDL or LDL cholesterol, where individual loci typically explain well below 1% of trait heritability [15,16]. This confirms the value of metabolites as useful intermediate endpoints for genetic studies of complex traits owing to their simpler allelic architecture [17].

To evaluate if our discovery effort based on HapMap2 panel comprehensively captured the effect of underlying causative genetic variants at the 145 loci, we explored the contribution of variants of lower allele frequency and of non-additive effects (epistasis) to metabolite heritability. To explore the contribution of low frequency variants, associations at each locus were recalculated after local imputation using a denser reference set (1000 Genomes Project, 1KGP), which allows more accurate imputation of rare variants compared to the sparser HapMap2 panel [18]. There was no improvement in variance explained using the 1KGP reference set (**Supplementary Table 7** and **Supplementary Table 8**, with the exception of the *CYP3A* cluster presented in **Supplementary Figure 3)**, suggesting that associations at these loci are well tagged by common variants well represented in the HapMap2 panel. We further systematically evaluated whether epistatic interactions between pairs of lead SNPs associated with the same metabolite may increase the proportion of heritability explained. This analysis did not suggest a major contribution for such epistatic effects to overall unexplained metabolite variance (**Supplementary Table 9**), with the exception of a strong interaction observed between *NAT8* and *PYROXD2* loci on the unknown metabolite X-12093 (**Figure 4**). Overall these observations suggest that associations at the 145 loci are explained by common genetic variants acting with predominantly additive effects, and that our discovery effort based on the HapMap2 panel was well powered to comprehensively capture the effect of these genetic variants.

eQTL annotation of metabolite loci and Mendelian randomization analysis

One major challenge of interpreting associations from GWAS is formulating and testing hypotheses on the causal effect of a SNP on an associated trait. One testable scenario on the effect of the SNP on the metabolite concentration (MET) is when it is mediated by gene expression (GE), which provides new opportunities to investigate metabolite pathways at the molecular level. SNP-GE-MET trios at a metabolite-associated SNP can thus be used to test whether a change in GE is causal to the relative change in MET, using SNP as instrumental variables. In this study, we exploited the availability of gene expression levels measured and at the same time of visit of the metabolomic measurement to explore causal effects at cis-regulatory SNPs.

To predict whether a lead SNP was likely to exert its effect on the metabolite through changes in expression of nearby genes, we systematically assessed the co-location of expression quantitative trait loci (eQTLs) with metabolic loci. Firstly, we assessed the eQTL overlap in two expression QTL datasets, and including: (i) the most recent iteration of the MuTHER eQTL dataset [19] and a liver eQTL dataset [20]. For each lead metabolomic SNP, we first retrieved all SNPs with high linkage disequilibrium ($r^2 \geq 0.8$) in the 1000 Genomes pilot phase (CEU population). Each lead SNP and its proxies were then used as baits to search the liver or MuTHER Project expression database. All significant *cis*-eQTLs within a 1Mb window centred on the lead SNP were retrieved from these dataset, and the best eQTL p-value in each tissue was noted. A total of 57 lead SNPs identified *cis*-eQTLs in at least one of four tissues searched under the nominal permutation p-value<0.001. A total of 101 SNP-gene pairs were identified, corresponding to 97 different genes. Of the 97 genes, 38 are annotated as causal to the metabolite association based on our annotation, and 59 as non-causal.

We first addressed the extent to which mQTLs overlap cis-eQTLs, and the extent to which such overlap can be interpreted in the context of possible tissue specific effects. Of the 38 causal genes, 10 (or 26% of total), 22 (58%), 17 (45%) and 17 (45%) had eQTLs in liver, fat, LCL and skin respectively (**Supplementary Table 10**). Of the 59 non-causal genes, 5 (8%), 21 (36%), 34 (58%), 27 (46%) had eQTLs in the same tissues. This suggests an enrichment by 3.25-fold of eQTLs matching causal genes in liver compared to those matching a non-causal gene (Fisher's exact test p-value = 0.023, two-tailed). Similarly, there was a 1.6-fold enrichment of eQTLs matching causal genes in fat compared to non-causal genes (Fisher's exact test p-value = 0.038, two-tailed). No enrichment was observed in LCL and skin, possibly reflecting the greater contribution liver and fat metabolism make to blood metabolite levels.

Having identified loci with a potential effect through gene regulation we next asked, for all 32 loci with eQTLs matching causal genes in fat, skin or LCLs, whether the change in GE was causal to the relative change in MET. For this analysis we exploited the unique availability of gene expression levels measured by the MuTHER resource in the same individuals and at the same time-point of metabolomics measurements [19]. Gene expression profiles for the 32 genes were retrieved from the MuTHER database in 484 TwinsUK participants overlapping with this study. We then applied a Mendelian randomization (MR) analysis to each SNP-transcript-metabolite dataset. Two loci showed significant evidence for gene expression changes mediating SNP-metabolite associations (*THEM4* and *CYP3A5,* **Figure 4**) under a Bonferroni-corrected permutation threshold accounting for the number of loci tested (p-value=$8.9\text{x}10^{-4}$=0.05/32). *THEM4* was significant also at a more stringent cutoff accounting for all 97 causal and non-causal genes associated with a SNP. Eight additional loci showed suggestive evidence at a nominal p-value of 0.05 (**Supplementary Table 11**). In these two cases, the Mendelian randomization formalization supports a causal role for the eQTL variants on metabolic trait associations, however in most cases the study power was not sufficient to conclusively demonstrate or refute causation.

The examples of *THEM4* and *CYP3A5* represent first examples of a formal evaluation of an assumption widely applied by many genome-wide association studies, i.e. that regulatory effects are likely to underpin associations with complex traits where overlap with eQTLs is observed. Our results thus provide a first paradigm for how genome-wide studies may in the future exploit such datasets to empower the downstream statistical evaluation of functional consequences of SNPs. The extension of this approach to larger datasets with similar molecular endpoints would allow systematically assessing causation at all putative cis-regulatory genes.

Network generation

Data preprocessing. To remove relatedness, we selected from each twin pair the individual with the least missing datapoints, leaving a total of 3,121 TwinsUK samples. We then sequentially excluded metabolites with more than 20% missing values, and samples with more than 10% missing values from the QC-ed metabolomics datasets for each cohort. Remaining missing values were imputed with the 'mice' package in R-project [23]. As a result, 3,047 unrelated TwinsUK samples with 355 metabolites and 1,764

KORA samples with 312 metabolites were left for the subsequent GGM network analysis.

Gaussian graphical modeling. We generated Gaussian graphical models (GGMs) for both metabolomics datasets as described previously [24]. Briefly, GGMs are based on partial correlation coefficients, i.e. pairwise correlations that have been corrected for the effects of all remaining variables in the dataset. Potential confounding effects from age, sex and batch effect were removed by adding them to the data matrix as well (thereby also correcting for these variables during partial correlation calculation). The full-order (conditioned against all other variables) partial correlation matrix $Z = \left(\zeta_{ij}\right)$ is defined as

$$\left(\zeta_{ij}\right) = -\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}} \text{ with } \left(\omega_{ij}\right) = \mathrm{P}^{-1},$$

where P is the matrix of regular Pearson correlation coefficients for all variables. Since the test for statistical significance of a partial correlation is heavily dependent on the respective sample sizes of the two studies, we here chose a constant positive partial correlation threshold of 0.2 in order to declare whether an edge is 'present' or not in the model. The use of a threshold based on partial correlation estimates, rather than the p-values of the partial correlations, was justified by the fact that the TwinsUK sample used for the GGM analysis has double the number of individuals than KORA. Because of the different statistical power of the two studies, the partial correlation value provides a more stable indicator of the partial correlation between two metabolites than the p-value itself, not affected by power. Supporting this choice, we consider that partial correlation estimates in cohorts with thousands of participants are very stable. We further removed negative partial correlations from this study, which may represent spurious signals. For instance, if A and B are uncorrelated, but both highly correlated with C, then the partial correlation between A and B will be strongly negative. Moreover, in a steady state mass network, substrates and products of an enzymatic reaction will not be negatively correlated (against common intuition). Both points have been discussed at greater length in a previous publication [24].

Combination of GGMs from the two cohorts. The two resulting GGMs from the KORA and the TwinsUK dataset were combined into a single consensus network by drawing an edge between two metabolites whenever at least one of the two models contains a significant partial correlation. The statistical difference between two partial correlations $\zeta_{\mathrm{ij}}^{\mathrm{A}}$ and $\zeta_{\mathrm{ij}}^{\mathrm{B}}$ between the same variables i and j from the two different datasets was assessed using the following statistic [25]:

$$\mathrm{T} = \frac{\mathrm{z}\left(\zeta_{\mathrm{ij}}^{\mathrm{A}}\right) - \mathrm{z}\left(\zeta_{\mathrm{ij}}^{\mathrm{B}}\right)}{\sqrt{1/(\mathrm{n_A} - \mathrm{p} - 1) + 1/(\mathrm{n_B} - \mathrm{p} - 1)}},$$

where $\mathrm{z}(\zeta) = \frac{1}{2}\ln\left(\frac{1+\zeta}{1-\zeta}\right)$ is the Fisher transformation [23], $\mathrm{n_A}$ and $\mathrm{n_B}$ are the sample sizes of datasets A and B, and p is the total number of variables in the two datasets (the union of variables going into GGM analysis, 396 metabolites). T is approximately standard normally distributed, and a statistical test can thus be constructed using the cumulative normal distribution function. Note that the partial correlation differences are annotated to the network edges in the online supplement.

Stability of the Gaussian graphical model. We performed a bootstrapping-based subsampling approach to verify the stability of partial correlations in the Twins and KORA datasets. To this end, we generated 1,000 bootstrap datasets for each cohort by randomly drawing from the original dataset with replacement. For each bootstrap dataset, we calculated the respective partial correlation matrix. The variation of partial correlation coefficients is generally low, indicating for stable estimates.

Combining metabolic and genetic networks. We added the results from our GWAS to the combined GGM network (see above). To this end, we introduced two new types of nodes, in addition to the metabolites: (1) Loci. A locus is linked with an edge to a metabolite if there is a genome-wide significant association between the metabolite concentration and at least one SNP in the locus region. (2) Ratios. In order to encode metabolite ratio information we introduced a group of ‘pseudo’-nodes. These nodes link two metabolites and one locus if the ratio of the two metabolites showed genome-wide significance with the locus. It is to be noted that generating a full joint graphical model including both metabolites and all SNPs is statistically not feasible due to the considerably large number of variables involved. Moreover, discrete variables like SNPs require more specialized types of graphical models than purely Gaussian ones (see e.g. [26]).

Condensed network generation. The ‘condensed’ network view (**Figure 2**) was generated as follows. Each metabolite is annotated with one out of the following eight super-pathways: “Lipid”, “Carbohydrate”, “Amino acid”, “Xenobiotics”, “Nucleotide”, “Energy”, “Peptide”, “Cofactors and vitamins”. The super-pathways are further subdivided into sub-pathways like “Oxidative phosphorylation”, “Carnitine metabolism”, or “Branched-chain amino acids”. To generate the network, we merged all metabolites belonging to the same sub-pathway into a single node. Two pathway-nodes were connected if there was at least one metabolite pair with a GGM edge, where one metabolite belongs to the one pathway and the other metabolite to the respective other pathway. Similarly, a pathway node and a locus were connected in the network, if there was at least one genome-wide significant association between the locus and one of the metabolites belonging to the pathway (or a ratio containing a metabolite from the pathway). Each pathway node was then colored according to its respective super-pathway, resulting in a total of eight different pathway colors in the final network.

The unknown metabolites in our analysis required a specific pre-processing for the network analysis. In order to incorporate the statistical associations with unknowns into the network, we derived their (most likely) sub-pathway annotation directly from the network context. To this end, we inspected the network neighborhood of the unknown metabolite. We assigned the major sub-pathway class among the known neighbors of the unknown node in the network (two nodes are neighbors if they share a common network edge). If all neighbors were either unknown metabolites themselves or gene locus nodes, we then investigated the 2-neighborhood (the neighbors of the neighbors), and so on. Unknown metabolites that could not be assigned any pathway annotation using this approach were excluded from the analysis. Note that the metabolomic and genetic network context has previously been previously shown to provide proper assignment of the metabolic pathway an unknown metabolite is involved in [6].

Integrating metabolic associations with complex trait locus information

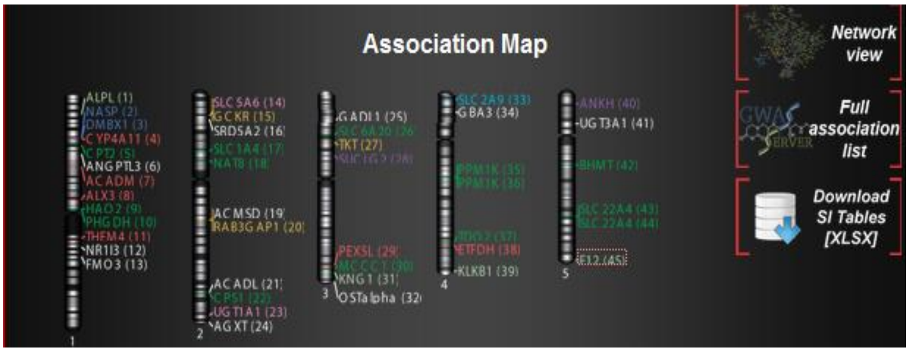
Supplementary Figure 5 provides an illustration of how one may integrate information from metabolic associations and relationships to increase understanding of complex trait associations. The bradykinin/kininogen/kinin system is a poorly understood system with a central role in the regulation of blood pressure. Our previous study reported an association between the kallikrein gene (*KLKB1*) and the des-Arg form of the nonapeptide bradykinin [5]. Bradykinin is a peptide hormone with a central role in the regulation of blood pressure central to the function of angiotensin-converting-enzyme (ACE) inhibitors. Two novel associations in this pathway were identified owing to the increased statistical power of our study. The first association was in *KNG1* (encoding kininogen 1), which undergoes alternative splicing to generate high molecular weight kininogen (HMWK- a precursor of bradykinin) and low molecular weight kininogen (LMWK). A second novel association was detected between variants in *F12* and the ratio between the unknown X-12038 and bradykinin. *F12* encodes Factor XII, a proenzyme activated to factor XIIa, responsible for the cleavage of prekallikrein (encoded by *KLKB1*) to form kallikrein by the cleavage of an internal Arg-Ile bond. Prekallikrein is a glycoprotein that participates in the surface-dependent activation of blood coagulation, fibrinolysis and inflammation, and contributes to feedback regulation of bradykinin levels [21]. These associations thus identify *KNG1* and *F12* variants at a center of an important cardiovascular disease pathway. Notably these variants were recently associated with activated partial thromboplastin time (APTT) [22], suggesting a potential role for these variants in the regulation of blood coagulation.

Outline and functionality of web resources

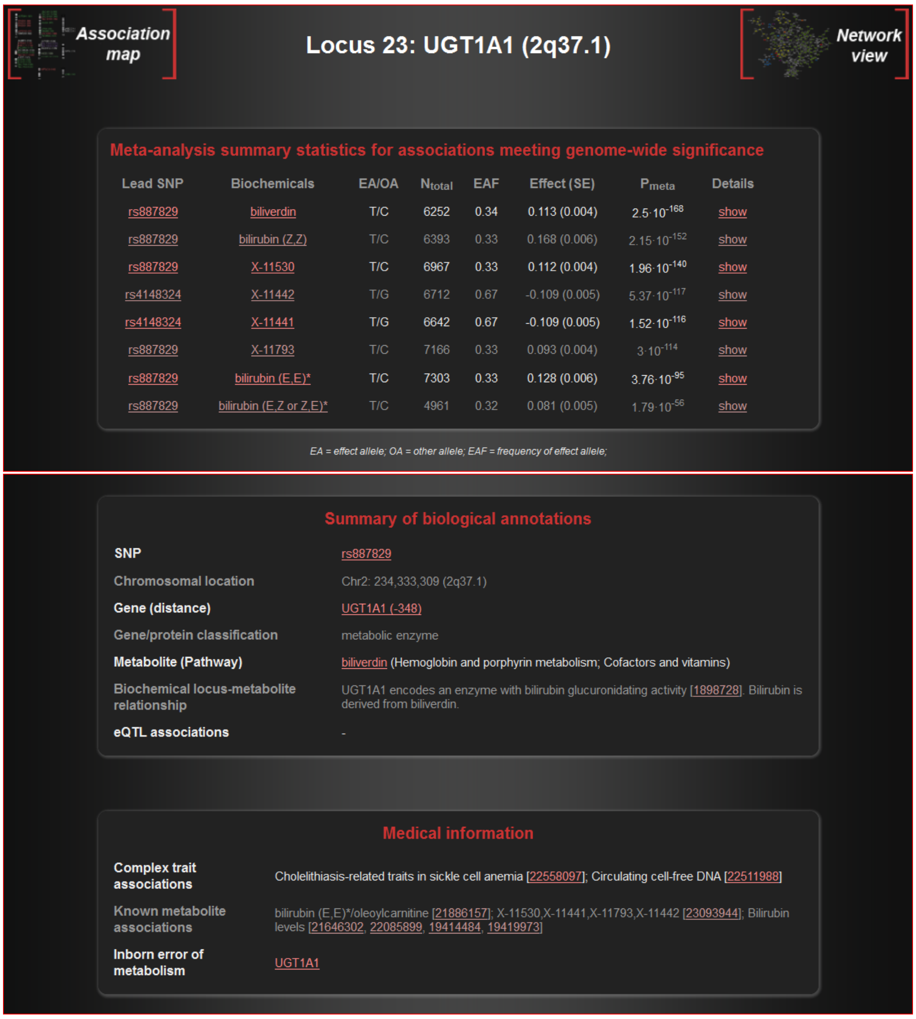
Supporting Online Website

To improve the accessibility of our data and results, we have developed an easy-to-use web resource (http://gwas.eu/si), which allows browsing and querying the data from various entry points:

1. *Chromosomal map* showing the positions of the 145 top loci serves as a gateway to all information collected for each genome-wide significant SNP-metabolite association.



2. *Locus information.* By clicking on a locus the user is provided with detailed information on the association, the extensive annotations regarding biological, medical and pharmaceutical relevance as well as further characteristics such as eQTL hits and metabolite heritability as well as locus-wide association plots. Metabolites and loci are linked to relevant public databases such as dbSNP, HMDB, KEGG, and orphanet.



Locus 23: UGT1A1 (2q37.1)

Meta-analysis summary statistics for associations meeting genome-wide significance

Lead SNP	Biochemicals	EA/OA	N_{total}	EAF	Effect (SE)	P_{meta}	Details
rs887829	biliverdin	T/C	6252	0.34	0.113 (0.004)	$2.5 \cdot 10^{-168}$	show
rs887829	bilirubin (Z,Z)	T/C	6393	0.33	0.168 (0.006)	$2.15 \cdot 10^{-152}$	show
rs887829	X-11530	T/C	6967	0.33	0.112 (0.004)	$1.96 \cdot 10^{-140}$	show
rs4148324	X-11442	T/G	6712	0.67	-0.109 (0.005)	$5.37 \cdot 10^{-117}$	show
rs4148324	X-11441	T/G	6642	0.67	-0.109 (0.005)	$1.52 \cdot 10^{-116}$	show
rs887829	X-11793	T/C	7166	0.33	0.093 (0.004)	$3 \cdot 10^{-114}$	show
rs887829	bilirubin (E,E)*	T/C	7303	0.33	0.128 (0.006)	$3.76 \cdot 10^{-95}$	show
rs887829	bilirubin (E,Z or Z,E)*	T/C	4961	0.32	0.081 (0.005)	$1.79 \cdot 10^{-56}$	show

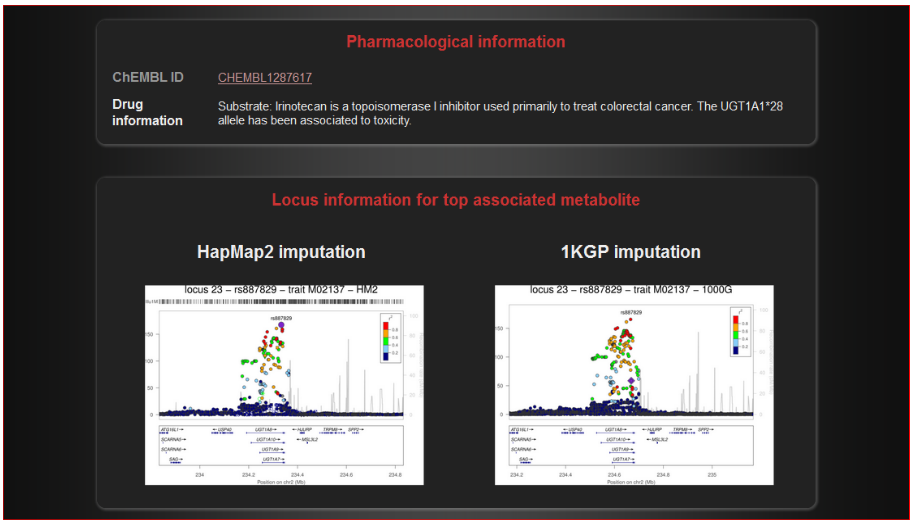
EA = effect allele; OA = other allele; EAF = frequency of effect allele;

Summary of biological annotations

SNP	rs887829
Chromosomal location	Chr2: 234,333,309 (2q37.1)
Gene (distance)	UGT1A1 (-348)
Gene/protein classification	metabolic enzyme
Metabolite (Pathway)	biliverdin (Hemoglobin and porphyrin metabolism; Cofactors and vitamins)
Biochemical locus-metabolite relationship	UGT1A1 encodes an enzyme with bilirubin glucuronidating activity [1898728]. Bilirubin is derived from biliverdin.
eQTL associations	-

Medical information

Complex trait associations	Cholelithiasis-related traits in sickle cell anemia [22558097]; Circulating cell-free DNA [22511988]
Known metabolite associations	bilirubin (E,E)*/oleoylcarnitine [21886157]; X-11530,X-11441,X-11793,X-11442 [23093944]; Bilirubin levels [21646302, 22085899, 19414484, 19419973]
Inborn error of metabolism	UGT1A1



Pharmacological information

ChEMBL ID	CHEMBL1287617
Drug information	Substrate: Irinotecan is a topoisomerase I inhibitor used primarily to treat colorectal cancer. The UGT1A1*28 allele has been associated to toxicity.

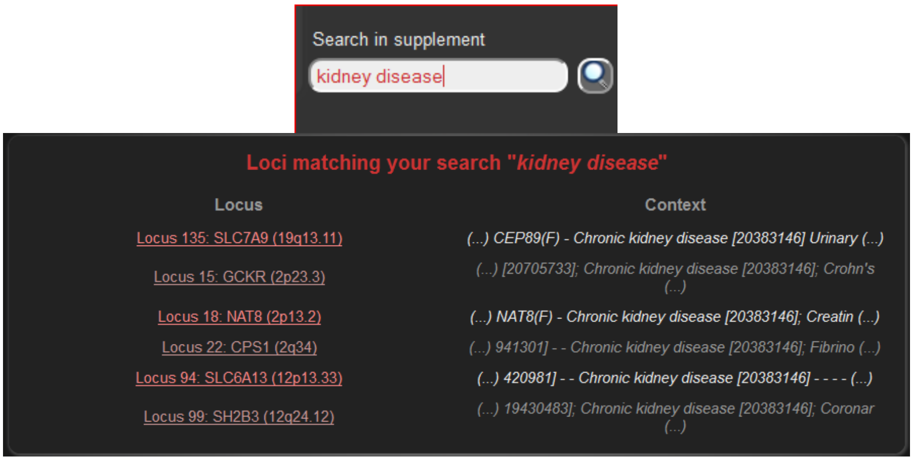
Locus information for top associated metabolite

HapMap2 imputation

1KGP imputation

locus 23 – rs887829 – trait M02137 – HM2
locus 23 – rs887829 – trait M02137 – 1000G
rs887829

3. *Free text search* in all information presented in the 145 locus web pages and given in the supplementary tables facilitates querying the data for readers with various levels of bioinformatics skills. Besides queries on rs numbers, gene and metabolite names, the system thus also allows queries such as and "peptide" or "kidney disease".



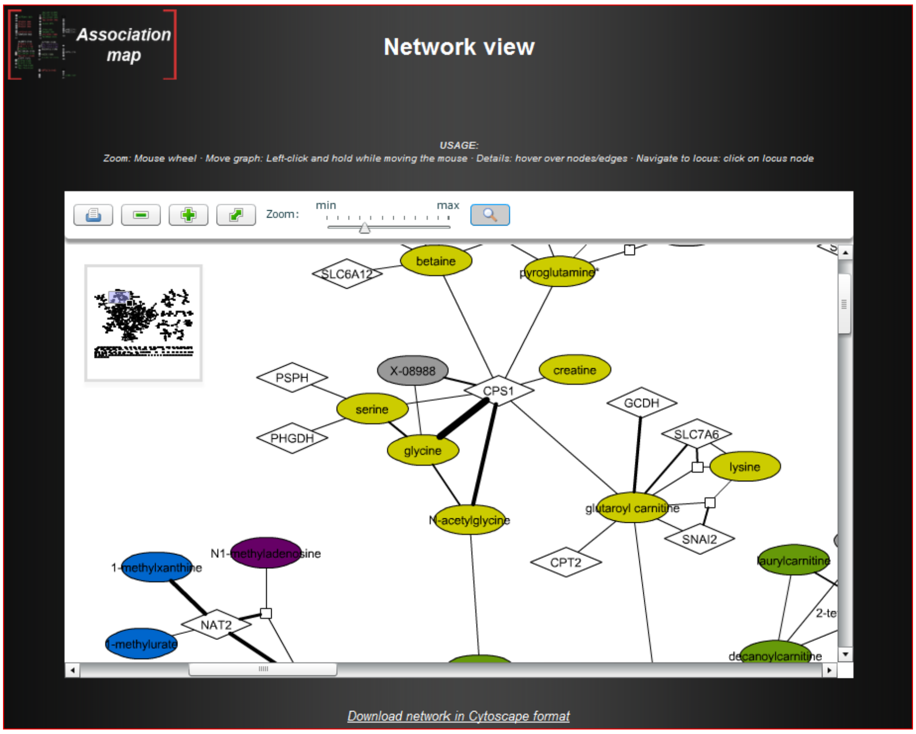
Search in supplement

kidney disease

Loci matching your search "*kidney disease*"

Locus	Context
Locus 135: SLC7A9 (19q13.11)	(...) CEP89(F) - Chronic kidney disease [20383146] Urinary (...)
Locus 15: GCKR (2p23.3)	(...) [20705733]; Chronic kidney disease [20383146]; Crohn's (...)
Locus 18: NAT8 (2p13.2)	(...) NAT8(F) - Chronic kidney disease [20383146]; Creatin (...)
Locus 22: CPS1 (2q34)	(...) 941301] - - Chronic kidney disease [20383146]; Fibrino (...)
Locus 94: SLC6A13 (12p13.33)	(...) 420981] - - Chronic kidney disease [20383146] - - - - (...)
Locus 99: SH2B3 (12q24.12)	(...) 19430483]; Chronic kidney disease [20383146]; Coronar (...)

4. *Web-based version of the complete reconstructed metabolic network* produced in this manuscript is accessible through direct browsing within the web site, while the corresponding stand-alone Cytoscape version is available from http://metabolomics.helmholtz-muenchen.de/gwa/si/network/SI_network.cys. Moreover, the network is interlinked with the supplement web pages for each locus. Via hyperlinks on these pages the reader can directly zoom into the part of the network relevant for the particular locus.



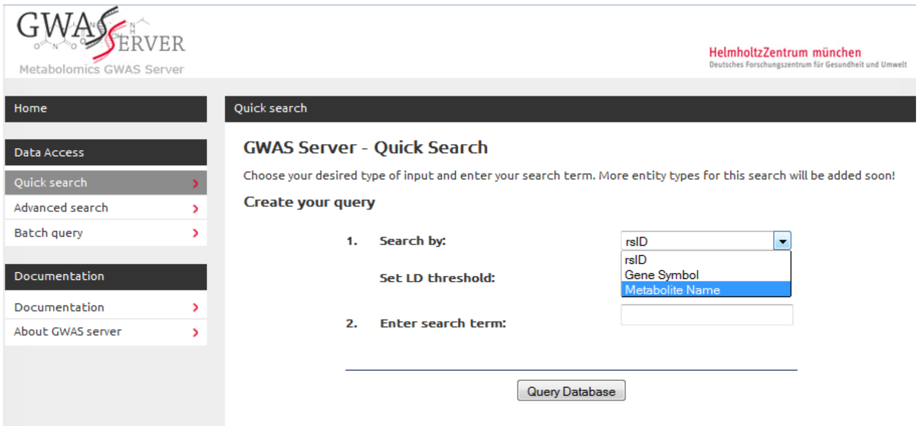
Network view

USAGE:
Zoom: Mouse wheel · Move graph: Left-click and hold while moving the mouse · Details: hover over nodes/edges · Navigate to locus: click on locus node

Download network in Cytoscape format

GWAS server

A GWAS server (http://metabolomics.helmholtz-muenchen.de/gwas) with efficient search functionality provides access to the full list of association results including associations that did not reach genome wide significance in this study but represent valuable information for researchers interested in specific genes or metabolites.



GWAS Server - Quick Search

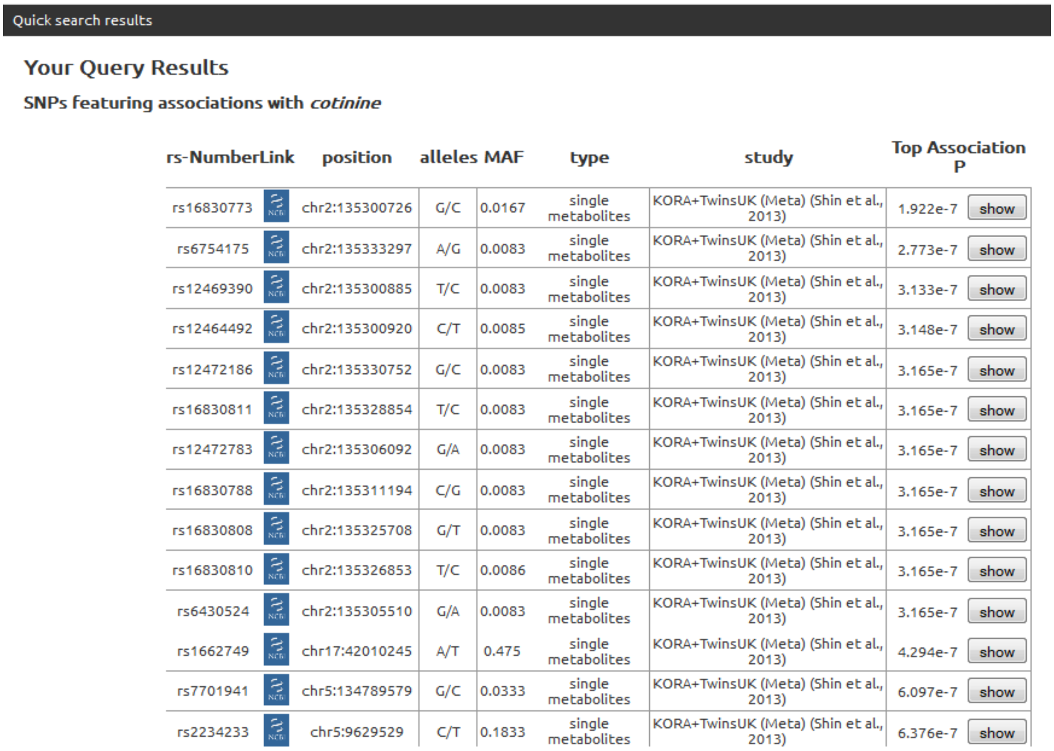
Choose your desired type of input and enter your search term. More entity types for this search will be added soon!

Create your query

1. Search by: rsID

Set LD threshold:

2. Enter search term:

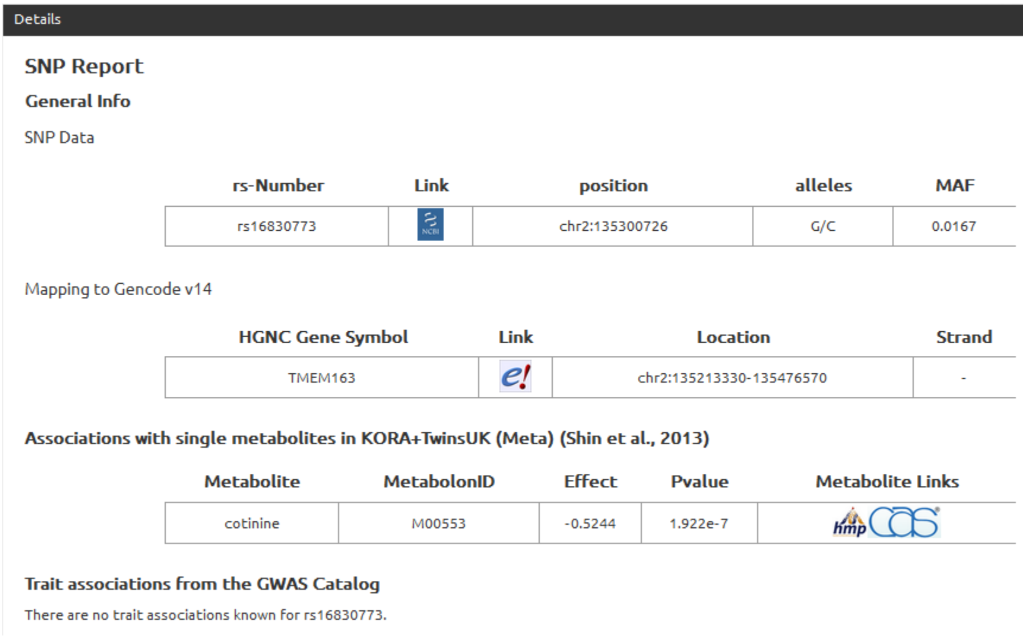


Quick search results

Your Query Results

SNPs featuring associations with *cotinine*

rs-NumberLink		position	alleles	MAF	type	study	Top Association P	
rs16830773	NCBI	chr2:135300726	G/C	0.0167	single metabolites	KORA+TwinsUK (Meta) (Shin et al., 2013)	1.922e-7	show
rs6754175	NCBI	chr2:135333297	A/G	0.0083	single metabolites	KORA+TwinsUK (Meta) (Shin et al., 2013)	2.773e-7	show
rs12469390	NCBI	chr2:135300885	T/C	0.0083	single metabolites	KORA+TwinsUK (Meta) (Shin et al., 2013)	3.133e-7	show
rs12464492	NCBI	chr2:135300920	C/T	0.0085	single metabolites	KORA+TwinsUK (Meta) (Shin et al., 2013)	3.148e-7	show
rs12472186	NCBI	chr2:135330752	G/C	0.0083	single metabolites	KORA+TwinsUK (Meta) (Shin et al., 2013)	3.165e-7	show
rs16830811	NCBI	chr2:135328854	T/C	0.0083	single metabolites	KORA+TwinsUK (Meta) (Shin et al., 2013)	3.165e-7	show
rs12472783	NCBI	chr2:135306092	G/A	0.0083	single metabolites	KORA+TwinsUK (Meta) (Shin et al., 2013)	3.165e-7	show
rs16830788	NCBI	chr2:135311194	C/G	0.0083	single metabolites	KORA+TwinsUK (Meta) (Shin et al., 2013)	3.165e-7	show
rs16830808	NCBI	chr2:135325708	G/T	0.0083	single metabolites	KORA+TwinsUK (Meta) (Shin et al., 2013)	3.165e-7	show
rs16830810	NCBI	chr2:135326853	T/C	0.0086	single metabolites	KORA+TwinsUK (Meta) (Shin et al., 2013)	3.165e-7	show
rs6430524	NCBI	chr2:135305510	G/A	0.0083	single metabolites	KORA+TwinsUK (Meta) (Shin et al., 2013)	3.165e-7	show
rs1662749	NCBI	chr17:42010245	A/T	0.475	single metabolites	KORA+TwinsUK (Meta) (Shin et al., 2013)	4.294e-7	show
rs7701941	NCBI	chr5:134789579	G/C	0.0333	single metabolites	KORA+TwinsUK (Meta) (Shin et al., 2013)	6.097e-7	show
rs2234233	NCBI	chr5:9629529	C/T	0.1833	single metabolites	KORA+TwinsUK (Meta) (Shin et al., 2013)	6.376e-7	show



Details

SNP Report

General Info

SNP Data

rs-Number	Link	position	alleles	MAF
rs16830773	NCBI	chr2:135300726	G/C	0.0167

Mapping to Gencode v14

HGNC Gene Symbol	Link	Location	Strand
TMEM163	e!	chr2:135213330-135476570	-

Associations with single metabolites in KORA+TwinsUK (Meta) (Shin et al., 2013)

Metabolite	MetabolonID	Effect	Pvalue	Metabolite Links
cotinine	M00553	-0.5244	1.922e-7	hmp CAS

Trait associations from the GWAS Catalog

There are no trait associations known for rs16830773.

Supplementary references

1. Krug, S. *et al.* The dynamic range of the human metabolome revealed by challenges. *FASEB J* (2012).
2. Evans, A.M., DeHaven, C.D., Barrett, T., Mitchell, M. & Milgram, E. Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Anal Chem* **81**, 6656-67 (2009).
3. Sha, W. *et al.* Metabolomic profiling can predict which humans will develop liver dysfunction when deprived of dietary choline. *FASEB J* **24**, 2962-75 (2010).
4. Dehaven, C.D., Evans, A.M., Dai, H. & Lawton, K.A. Organization of GC/MS and LC/MS metabolomics data into chemical libraries. *J Cheminform* **2**, 9 (2010).
5. Suhre, K. *et al.* Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477**, 54-60 (2011).
6. Krumsiek, J. *et al.* Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. *PLoS Genet* **8**, e1003005 (2012).
7. Horai, H. *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* **45**, 703-14 (2010).
8. Sreekumar, A. *et al.* Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* **457**, 910-914 (2009).
9. Illig, T. *et al.* A genome-wide perspective of genetic variation in human metabolism. *Nature Genetics* **42**, 137-141 (2010).
10. Soranzo, N. *et al.* Meta-analysis of genome-wide scans for human adult stature identifies novel Loci and associations with measures of skeletal frame size. *PLoS Genetics* **5**, e1000445 (2009).
11. Kolz, M. *et al.* Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations. *PLoS Genetics* **5**, e1000504 (2009).
12. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics* **5**, e1000529 (2009).
13. Petersen, A.K. *et al.* On the hypothesis-free testing of metabolite ratios in genome-wide and metabolome-wide association studies. *BMC Bioinformatics* **13**, 120 (2012).
14. Boker, S. *et al.* OpenMx: An Open Source Extended Structural Equation Modeling Framework. *Psychometrika* **76**, 306-317 (2011).
15. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832-838 (2010).
16. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-713 (2010).
17. Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature Genetics* **44**, 269-276 (2012).
18. Huang, J., Ellinghaus, D., Franke, A., Howie, B. & Li, Y. 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. *Eur J Hum Genet* **20**, 801-5 (2012).
19. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics* **44**, 1084-1089 (2012).
20. Schadt, E.E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* **6**, e107 (2008).
21. Chung, D.W., Fujikawa, K., McMullen, B.A. & Davie, E.W. Human plasma prekallikrein, a zymogen to a serine protease that contains four tandem repeats. *Biochemistry* **25**, 2410-7 (1986).
22. Houlihan, L.M. *et al.* Common variants of large effect in F12, KNG1, and HRG are associated with activated partial thromboplastin time. *Am J Hum Genet* **86**, 626-31 (2010).
23. Buuren, S.v. & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* **45**, 1-67 (2011).
24. Krumsiek, J., Suhre, K., Illig, T., Adamski, J. & Theis, F.J. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst Biol* **5**, 21 (2011).
25. Levy, K.J. & Narula., S.C. Testing Hypotheses concerning Partial Correlations: Some Methods and Discussion. *International Statistical Review* **46**, 215-218 (1978).

26. Fellinghauer, B., Bühlmann, P., Ryffel, M., von Rhein, M. & Reinhardt, J.D. Stable graphical model estimation with Random Forests for discrete, continuous, and mixed variables. *Computational Statistics & Data Analysis* **64**, 132-152 (2013).

27. Sainz, I.M., Pixley, R.A. & Colman, R.W. Fifty years of research on the plasma kallikrein-kinin system: from protein structure and function to cell biology and in-vivo pathophysiology. *Thromb Haemost* **98**, 77-83 (2007).

MuTHER consortium

Kourosh R. Ahmadi[1], Chrysanthi Ainali[2], Amy Barrett[3], Veronique Bataille[1], Jordana T. Bell1[4], Alfonso Buil[5], Panos Deloukas[6], Emmanouil T. Dermitzakis[5], Antigone S. Dimas[4,5], Richard Durbin[6], Daniel Glass[1], Elin Grundberg[1,6], Neelam Hassanali[3], Åsa K. Hedman[4], Catherine Ingle[6], David Knowles[7], Maria Krestyaninova[8], Cecilia M. Lindgren[4], Christopher E. Lowe[9,10], Mark I. McCarthy[3,4,11], Eshwar Meduri[1,6], Paola di Meglio[12], Josine L. Min[4], Stephen B. Montgomery[5], Frank O. Nestle[12], Alexandra C. Nica[5], James Nisbet[6], Stephen O'Rahilly[9,10], Leopold Parts[6], Simon Potter[6], Magdalena Sekowska[6], So-Youn Shin[6], Kerrin S. Small1[6], Nicole Soranzo[6], Tim D. Spector[1], Gabriela Surdulescu[1], Mary E. Travers[3], Loukia Tsaprouni[6], Sophia Tsoka[2], Alicja Wilk[6], Tsun-Po Yang[6], Krina T. Zondervan[4]

[1] Department of Twin Research and Genetic Epidemiology, King's College London, London, UK
[2] Department of Informatics, School of Natural and Mathematical Sciences, King's College London, Strand, London, UK
[3] Oxford Centre for Diabetes, Endocrinology & Metabolism, University of Oxford, Churchill Hospital, Oxford, UK
[4] Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
[5] Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland
[6] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK
[7] University of Cambridge, Cambridge, UK
[8] European Bioinformatics Institute, Hinxton, UK
[9] University of Cambridge Metabolic Research Labs, Institute of Metabolic Science Addenbrooke's Hospital Cambridge, UK
[10] Cambridge NIHR Biomedical Research Centre, Addenbrooke's Hospital, Cambridge, UK
[11] Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, UK
St. John's Institute of Dermatology, King's College London, London, UK

Supplementary Tables and Figures

Supplementary Table 1. Study descriptives

Details of methods and software used for the genetic analysis, including cohort samples, metrics used for genotyping, SNP data quality control and imputation and details on statistical analyses.

Supplementary Table 2. Metabolites descriptives

Pathway and super-pathway information are given for metabolites with known chemical identity. Information on measurement platform, number of individuals with valid trait measurement as well as trait mean, SD and range after QC (composed of run-day block correction also known as run-day normalization, log transformation with a base of 10 and >4SD outliers exclusion) are given for all metabolites. The spectra data format (129:15) indicates the ion (m/z) and relative peak intensity. Each ion in the spectra is separated by a space. For CV(%) calculation, 4-5 replicates of the MTRX (QC/technical replicate samples created from a pool of well characterized human plasma) were run each platform day, corresponding to 1300 MTRX for the study.

Supplementary Table 3. Correlation between metabolites

Pearson's correlation coefficient r between pairs of metabolites measured in this study.

Supplementary Table 4. Summary statistics for the 145 loci identified in this study

Only the most associated metabolite per locus is given. Genes with a high plausibility of being causal based on known biochemical function (Online Methods) are underlined; for other loci, the gene nearest to the sentinel SNP is given. A full list of all metabolites associated at genome-wide significance with each locus, and study specific summary statistics, are given in **Supplementary Table 5**. Associations below genome-wide significance are available from the Metabolomics GWAS server (see **URLs**).

Supplementary Table 5. Study-specific association statistics

Statistics are reported for the most associated metabolite at each locus in **Supplementary Table 4**. EA/OA = effect/other allele; EAF = effect allele frequency.

Supplementary Table 6. Summary of biologic and disease annotations

For each of the 145 loci, information correlating metabolites to gene functions is given, indicating likely biochemical reactions underlying the observed associations. Genes within 500 kb from the lead SNP were annotated as detailed in the **Online Methods**. GWAS overlap was annotated by searching the complete NHGRI GWAS database using either the lead SNP or a statistical equivalent (proxy, with $r^2 \geq 0.8$ with lead SNP). Drug information: ME=Metabolizing enzyme, DT=drug target or TP=transporter for FDA- and/or EMA-approved drug; Dev=in development, Drug-like=compound with activity in ChEMBL).

Supplementary Table 7. Heritability of metabolites and variance explained

For each of 486 raw metabolites, the narrow-sense heritability was estimated under a full ACE model from monozygotic and dizygotic twin pairs in TwinsUK. The known heritability (or the variance explained by known variants in GWAS) was estimated by multiple regression analysis including all associated SNPs under additive genetic models, after adjusting for covariates (age and sex in KORA; age, sex and experimental batch in TwinsUK) both in TwinsUK (limited to a subset of unrelated singletons) and KORA. The variance explained by SNPs was estimated using the most associated SNPs identified based on HapMap2 and 1000 Genome Project analyses respectively (**Supplementary Table 8**).

Supplementary Table 8. Associations based on 1000 Genomes Project

For each locus described in **Supplementary Table 4**, the identity of the most associated SNP was obtained from the discovery effort (based on HapMap2 imputation) to the best SNP identified after local re-analysis based on imputation from the 1000 Genome Project. The variances explained by the most associated SNP in both HM2 and 1KG based genotype data are given as well as effect sizes, standard errors and p-values. The variance explained by each SNP in TwinsUK was estimated using a subset of unrelated samples.

Supplementary Table 9. Epistatic effects

Additive and interaction models were fit for each pairs of SNPs associated with a given metabolite at genome-wide significant level (additive: Y = alpha + beta1*SNP1 + beta2*SNP2 + e; interaction: Y = alpha + beta1*SNP1 + beta2*SNP2 + gamma*SNP1*SNP2 +e). Significance of the interaction term in the epistasis model was tested by ANOVA F-test.

Supplementary Table 10. Loci overlapping cis-eQTLs in four tissues

Overlap of metabolomic loci with expression quantitative trait loci (*cis*-eQTL) measured in Liver [20] or in fat, skin or LCLs by the MuTHER project [19]. For each lead metabolomic SNP, we first retrieved all SNPs with high linkage disequilibrium ($r^2 \geq 0.8$) in the 1000 Genomes pilot phase (CEU population). Each lead SNP and its proxies were then used as baits to search the MuTHER Project expression database. All significant *cis*-eQTLs within a 1Mb window centred on the lead SNP were retrieved from these dataset, and the best eQTL p-value in each tissue was noted. All eQTLs identified in the MuTHER dataset and where the gene matched the gene prioritized as likely causal, were taken forward in the Mendelian randomization analysis (**Supplementary Table 11**).

Supplementary Table 11. Exploration of causality at SNP-GE-metabolite trios

For each of the 32 loci where the lead SNP (or its proxy) was *cis*-regulatory in the MuTHER Pilot database and matched a predicted causal gene, the effect of the probe gene expression (GE) onto the metabolite (MET) was estimated by Mendelian randomization using a subset of 484 unrelated twin with gene expression and metabolite measurements taken at the same time of visit. A Bonferroni corrected 99.85% confidence interval (p-value=0.05/32) was obtained from 10,000 permutations.

Supplementary Table 12. Inborn errors of metabolism

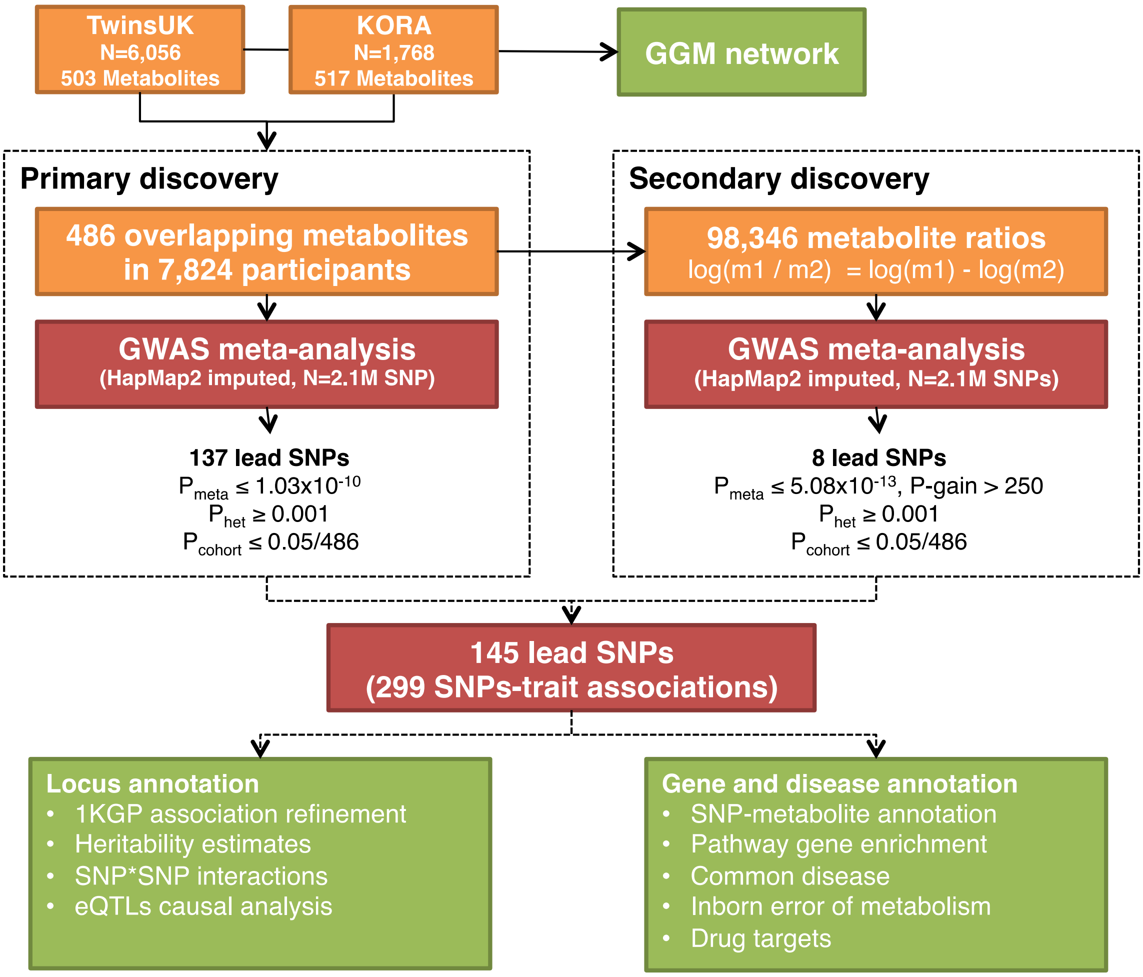
Information on gene and disease characteristics for genes associated with inborn errors of metabolism and overlapping with the metabolomic loci were obtained from Orphanet.

Supplementary Table 13. Drug targets and corresponding drugs

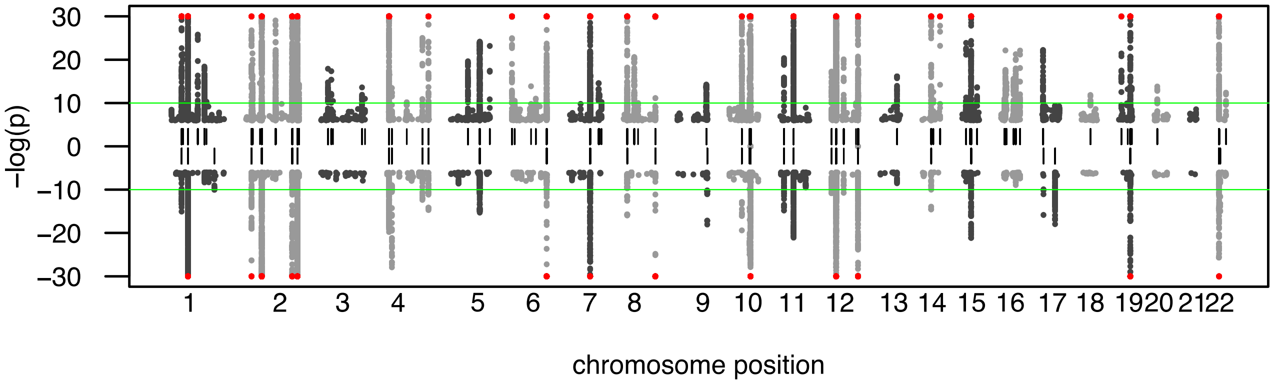
List of FDA-approved drugs that exert their therapeutic effect through genes and proteins at each locus.

Supplementary Table 14. Summary of targets for drugs in different stages of development

Information is presented only for targets for drugs in various stages of development (preclinical, Phase I-III, pre-registration or registration), failed (either discontinued, withdrawn) or where no development activity is reported (NDR) in the PharmaProjects resource.

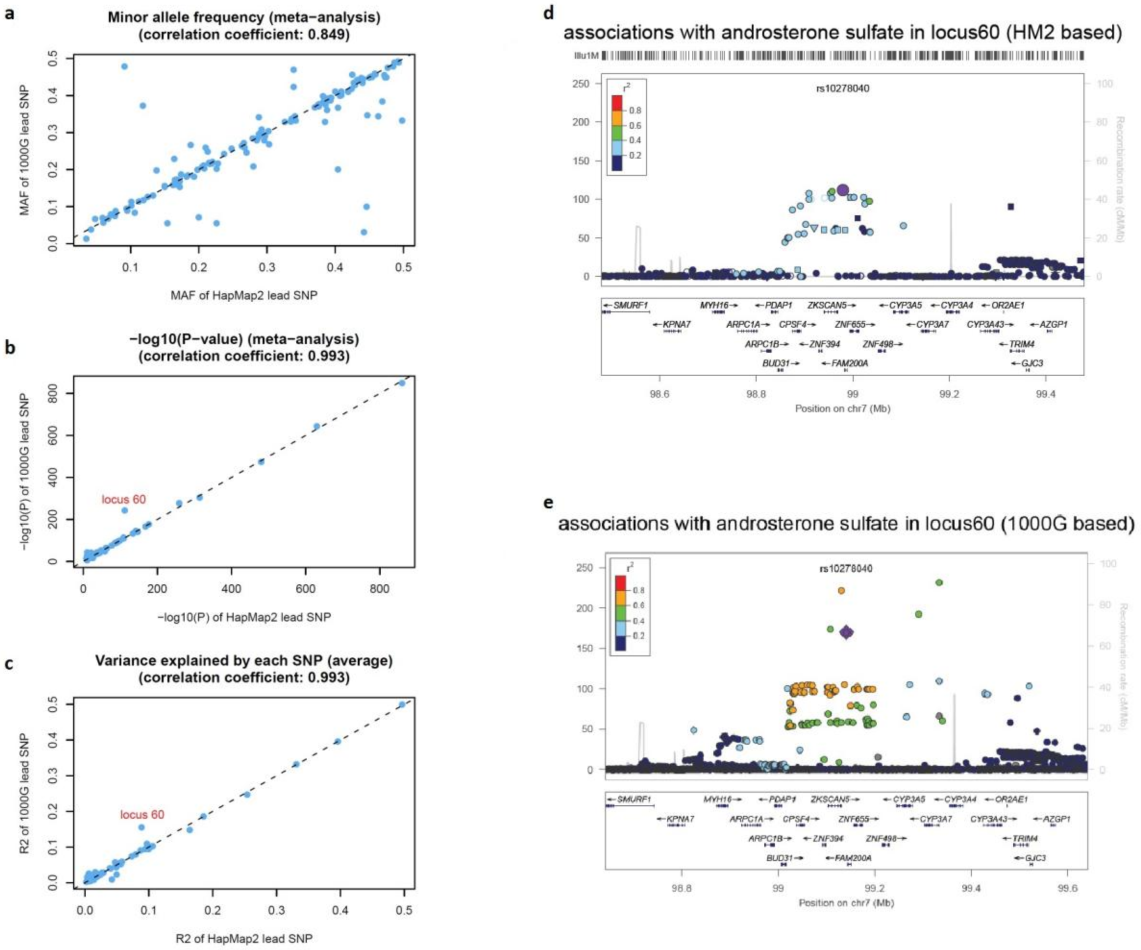

TwinsUK
N=6,056
503 Metabolites
KORA
N=1,768
517 Metabolites
GGM network
Primary discovery
486 overlapping metabolites in 7,824 participants
GWAS meta-analysis
(HapMap2 imputed, N=2.1M SNP)
137 lead SNPs
$P_{meta} \leq 1.03x10^{-10}$
$P_{het} \geq 0.001$
$P_{cohort} \leq 0.05/486$
Secondary discovery
98,346 metabolite ratios
log(m1 / m2) = log(m1) - log(m2)
GWAS meta-analysis
(HapMap2 imputed, N=2.1M SNPs)
8 lead SNPs
$P_{meta} \leq 5.08x10^{-13}$, P-gain > 250
$P_{het} \geq 0.001$
$P_{cohort} \leq 0.05/486$
145 lead SNPs
(299 SNPs-trait associations)
Locus annotation
• 1KGP association refinement
• Heritability estimates
• SNP*SNP interactions
• eQTLs causal analysis
Gene and disease annotation
• SNP-metabolite annotation
• Pathway gene enrichment
• Common disease
• Inborn error of metabolism
• Drug targets

Supplementary Figure 1. Study design



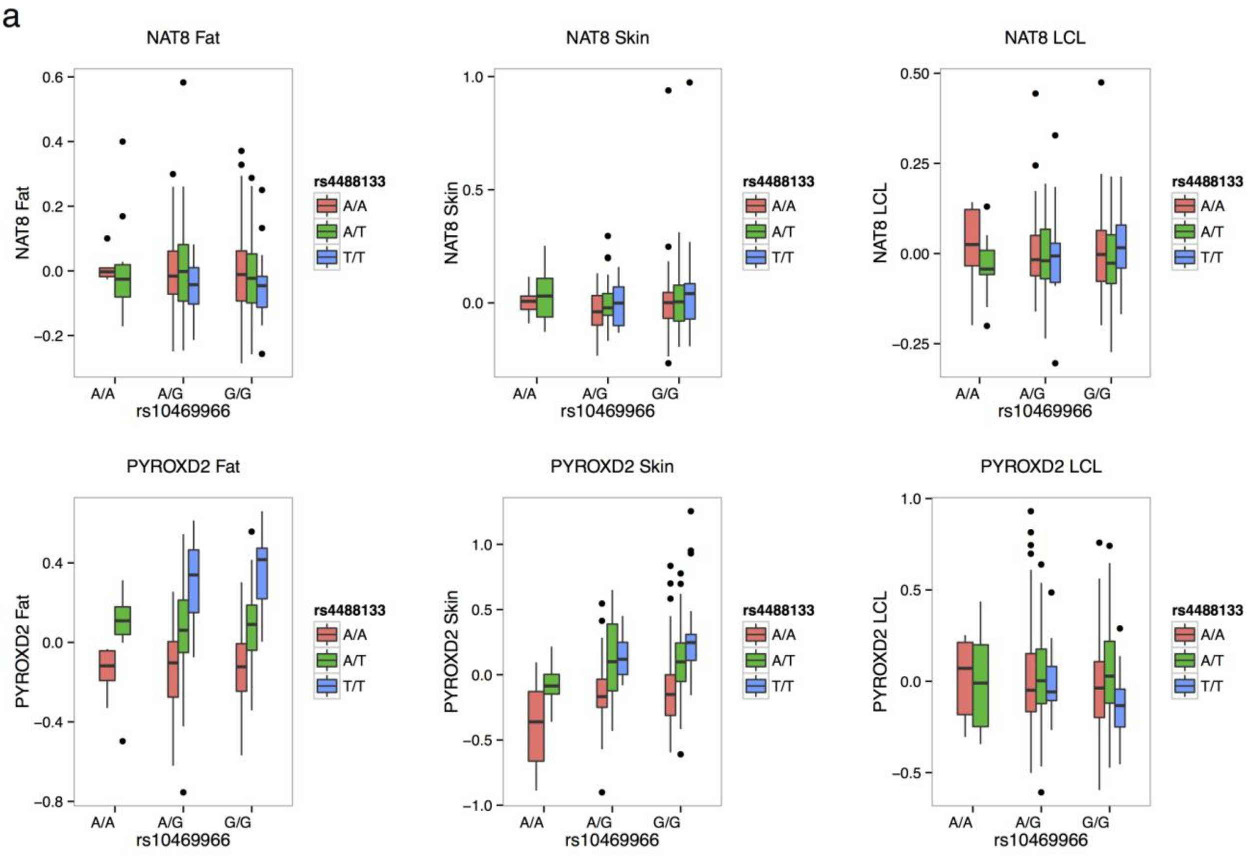
Supplementary Figure 2. Manhattan plots

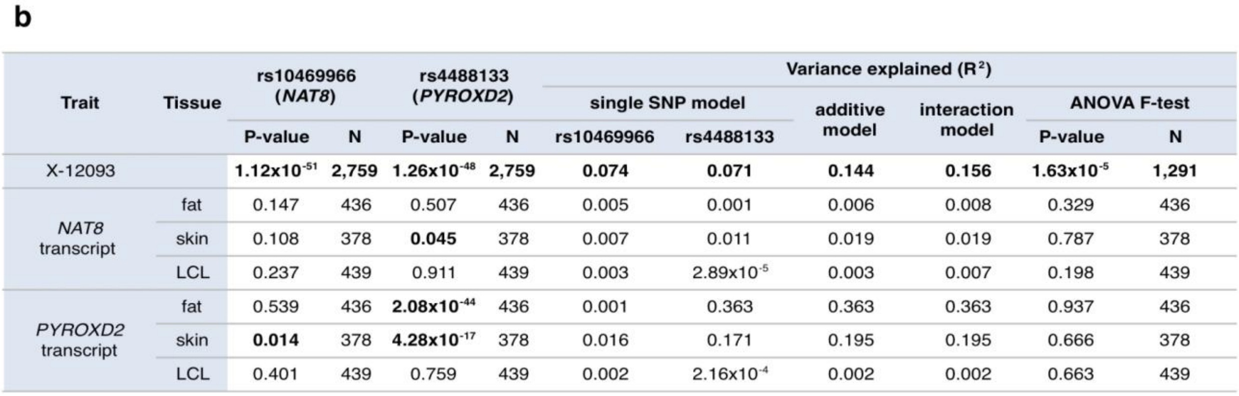
Association results for raw metabolite concentrations are shown for genome-wide SNPs. Top panel: TwinsUK, bottom: KORA. Only SNPs with $p<1x10^{-6}$ are displayed. The green line indicates the genome-wide cut-off of $p<1.03x10^{-10}$. Loci with P-values $<1x10^{-30}$ are indicated with a red symbol.



Supplementary Figure 3. Comparison of imputations based on HapMap2 and 1000 Genomes Project

Correlation between **a)** minor allele frequency (MAF) in meta-analysis, **b)**, association p-value in meta-analysis (on a –log10 scale), **c)** and average variance explained for the most significant SNPs selected from imputation based on either the HM2 (x-axis) or 1KG (y-axis) panels. The high correlations between the HapMap2 and 1KGP datasets support the view that metabolic associations are driven by common variants well tagged by HM2 imputation. **d, e**) One exception is locus *CYP3A4-5-7*, where the 1KGP scan reveals an additional variant (rs10278040) with greater association and variance explained for androsterone sulfate compared to the corresponding HM2 variant (rs148982377). [HM2: rs148982377, MAF=0.038, P=7.65x10^{-244}, R^2=15.6%; 1KGP: rs10278040, MAF=0.042, P=8.82x10^{-113}, R^2_{HM2}=10.3%].



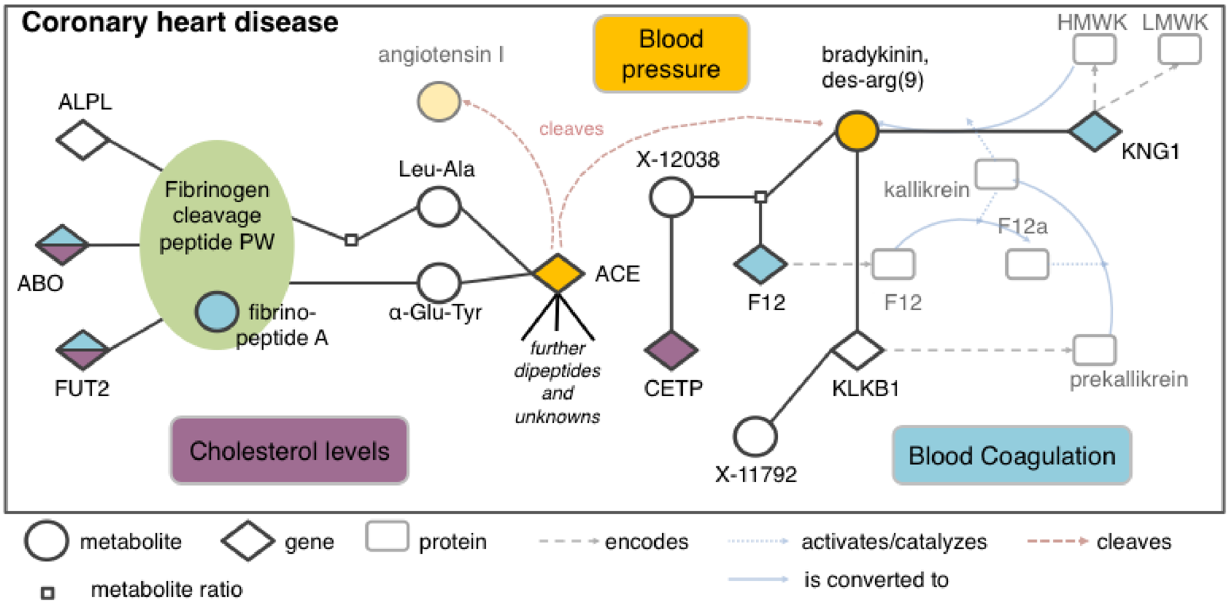


Trait	Tissue	rs10469966 (*NAT8*)		rs4488133 (*PYROXD2*)		Variance explained (R^2)					
						single SNP model		additive model	interaction model	ANOVA F-test	
		P-value	N	P-value	N	rs10469966	rs4488133			P-value	N
X-12093		**1.12×10^{-51}**	**2,759**	**1.26×10^{-48}**	**2,759**	**0.074**	**0.071**	**0.144**	**0.156**	**1.63×10^{-5}**	**1,291**
NAT8 transcript	fat	0.147	436	0.507	436	0.005	0.001	0.006	0.008	0.329	436
	skin	0.108	378	**0.045**	378	0.007	0.011	0.019	0.019	0.787	378
	LCL	0.237	439	0.911	439	0.003	2.89×10^{-5}	0.003	0.007	0.198	439
PYROXD2 transcript	fat	0.539	436	**2.08×10^{-44}**	436	0.001	0.363	0.363	0.363	0.937	436
	skin	**0.014**	378	**4.28×10^{-17}**	378	0.016	0.171	0.195	0.195	0.666	378
	LCL	0.401	439	0.759	439	0.002	2.16×10^{-4}	0.002	0.002	0.663	439

Supplementary Figure 4. Interaction between *NAT8* and *PYROXD2* variants

a) Boxplots of *PYROXD2* and *NAT8* transcript levels in fat, skin and LCLs as a function of the genotype conformation between the two variants rs10469966 (*NAT8)* and rs4488133 (*PYROXD2)*.

b) Summary of association and interaction effects at the two loci, summarizing association statistics and variances explained under the single SNP, additive and interaction models. An ANOVA F-test was used to test significance of the interactive model over the additive model. The association test with X-12093 reported here is based on combined TwinsUK and KORA dataset; all other analyses were carried out using unrelated TwinsUK singletons. See also **Figure 4**.



Supplementary Figure 5. Cardiovascular disease and hypertension metabolic sub-network

Network data was annotated with expert knowledge to illustrate correlations between molecular relationships and knowledge on blood pressure regulation, blood coagulation, and known molecular risk factors for cardiovascular disease and hypertension. **Black nodes and edges.** This sub-network was derived from metabolite data and corresponds to the inset in **Figure 2 B.** Metabolites (circular nodes) and genes (diamond-shaped nodes) of the fibrinogen cleavage (left) and the kininogen/kinin system (right) and their interconnections were derived from our data. **Grey nodes and edges.** Annotations of biochemical function based on expert knowledge [27]. **Colored nodes and edges**. Reported associations based on genome-wide studies for blood pressure regulation (orange), blood coagulation (blue), and cholesterol levels (purple; information in **Supplementary Table 6**).